PLOS ONE

RESEARCH ARTICLE

Annotation, classification, genomic organization and expression of the *Vitis vinifera* CYPome

Tina Ilc[1], Gautier Arista[2], Raquel Tavares[3], Nicolas Navrot[1], Eric Duchêne[2], Amandine Velt[2], Frédéric Choulet[4], Etienne Paux[4], Marc Fischer[2], David R. Nelson[5], Philippe Hugueney[2], Danièle Werck-Reichhart[1], Camille Rustenholz[2]*

1 Institute of Plant Molecular Biology, Centre National de la Recherche Scientifique, Université de Strasbourg, Strasbourg, France, **2** Université de Strasbourg, INRA, SVQV UMR-A 1131, Colmar, France, **3** Laboratoire de Biométrie et Biologie Evolutive, Centre National de la Recherche Scientifique, Université de Lyon 1, Lyon, France, **4** Laboratoire Structure et Evolution du Génome du Blé, Institut National de la Recherche Agronomique, Université Blaise Pascal, Clermont-Ferrand, France, **5** Department of Microbiology, Immunology and Biochemistry, University of Tennessee Health Science Center, Memphis, Tennessee, United States of America

* camille.rustenholz@inra.fr

Citation: Ilc T, Arista G, Tavares R, Navrot N, Duchêne E, Velt A, et al. (2018) Annotation, classification, genomic organization and expression of the *Vitis vinifera* CYPome. PLoS ONE 13(6): e0199902. https://doi.org/10.1371/journal.pone.0199902

Editor: Sara Amancio, Universidade de Lisboa Instituto Superior de Agronomia, PORTUGAL

Received: December 20, 2017

Accepted: June 15, 2018

Published: June 28, 2018

Data Availability Statement: RNA-Seq datasets of green and mid-ripening berries from Riesling and Gewurztraminer were submitted at NCBI SRA public database under the BioProject accession number PRJNA378596.

Funding: The authors acknowledge the support of the French Agence Nationale de la Recherche to the InteGrape project (ANR-13-BSV6-0010). The doctoral fellowship of Tina Ilc was funded by the People Programme (Marie Curie Actions) of the European Union's 7th Framework Programme

Abstract

Cytochromes P450 are enzymes that participate in a wide range of functions in plants, from hormonal signaling and biosynthesis of structural polymers, to defense or communication with other organisms. They represent one of the largest gene/protein families in the plant kingdom. The manual annotation of cytochrome P450 genes in the genome of *Vitis vinifera* PN40024 revealed 579 P450 sequences, including 279 complete genes. Most of the P450 sequences in grapevine genome are organized in physical clusters, resulting from tandem or segmental duplications. Although most of these clusters are small (2 to 35, median = 3), some P450 families, such as CYP76 and CYP82, underwent multiple duplications and form large clusters of homologous sequences. Analysis of gene expression revealed highly specific expression patterns, which are often the same within the genes in large physical clusters. Some of these genes are induced upon biotic stress, which points to their role in plant defense, whereas others are specifically activated during grape berry ripening and might be responsible for the production of berry-specific metabolites, such as aroma compounds. Our work provides an exhaustive and robust annotation including clear identification, structural organization, evolutionary dynamics and expression patterns for the grapevine cytochrome P450 families, paving the way to efficient functional characterization of genes involved in grapevine defense pathways and aroma biosynthesis.

Introduction

Grapevine (*Vitis vinifera L.*) is one of the oldest [1] and economically the most important [2] fruit crop in the world. The majority of grapes produced worldwide are used in winemaking. Modern cultivated grapevine has been shaped by thousands of years of selection for traits such

(FP7/2007-2013) under REA Grant Agreement 289217. The doctoral fellowship of Gautier Arista was funded by INRA and the Région Alsace. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

as berry size, sugar content or skin color [3], but today's viticulture is facing new challenges. In addition to pathogen pressure, it has to deal with climate change [4,5], and shift of consumer preference towards higher quality wines with a lower environmental impact [6,7]. Traditional breeding is extremely difficult to apply in grapevine because of its long lifecycle, reduced fitness of progeny and complexity of quality traits [8]. Sequencing of the grapevine genome in 2007 [9] and advances in the 'omics' techniques [10] set the stage for more efficient breeding solutions. The next crucial step towards improved grapevine varieties is the identification of genes underlying important traits, such as response towards pathogens, fruit development and quality.

Many developmental as well as ecological response pathways in plants involve cytochrome P450 oxygenases [11,12]. In plants, these enzymes catalyze regio- and stereospecific insertion of an oxygen atom into small, hydrophobic substrates that range from terpenoids and fatty acids to amino acids and their derivatives, such as phenolic compounds. In the model plant *Arabidopsis thaliana* they control processes as diverse as plant growth and branching [13,14], flower [15,16] and fruit development [17], formation of lignin and surface biopolymers [18,19], emission of volatiles [20,21] or plant-pathogen and plant-insect interactions [21–23]. In crop plants, P450s play major roles in shaping agriculturally-relevant traits, such as fruit size [24] or aroma biosynthesis [25]. This makes cytochromes P450 attractive targets for crop improvement.

Cytochromes P450 in plants evolved into many distinct families, which are usually composed by genes with 40% or higher protein sequence identity. Within one P450 family the biochemical function is often conserved across the plant kingdom. For example, enzymes from the CYP97 family are involved in carotenoid hydroxylation, CYP79s in the *N*-hydroxylation of amino acid to aldoximes, CYP75s in the hydroxylation of flavonoids, and CYP704s in addition to CYP703s in fatty acid hydroxylation to form the precursors to structural polymers sporopollenin and cutin [26]. Members of other families, however, have divergent functions: some members of CYP72 family are involved in iridoid biosynthesis, whereas others oxidize triterpene substrates [27]. These differences stem from different evolutionary pressures on genes with different functions. Families with essential functions, such as hormone metabolism or synthesis of biopolymers, usually show a low copy number and are submitted to high purifying selection, whereas families with adaptive functions expanded or "bloomed" in certain taxa [28]. A well-documented example is the bloom of the CYP76M subfamily in rice (*Oryza sativa*), which consists of 11 genes and 2 pseudogenes. At least 4 members of this subfamily are involved in the biosynthesis of diterpenoid antifungal compounds [29,30]. They are clustered close together in the genome, which is another common feature of recently duplicated P450s and probably result from sequential tandem duplications [28]. Interestingly, in other plants, for example *Arabidopsis thaliana* or *Catharanthus roseus*, some CYP76 members have a different biochemical function, namely oxidation of monoterpenols or their iridoid derivatives [31,32]. Recently expanded P450 families might therefore have new ecological functions, but those are more difficult to predict compared to functions of conserved P450 families. In addition, function of many P450 families is still unknown or poorly understood.

A previous annotation of P450s has highlighted some potentially interesting gene families in the highly heterozygous *V. vinifera* cv. Pinot Noir genome [33–35]. In this work we performed the first complete manual annotation of P450s in the nearly homozygous *V. vinifera* reference genome PN40024 [9]. We discuss the structural organization of the genes with particular focus on gene clusters. We evaluate phylogenetic relationships between those genes in order to be able to identify recently expanded gene families likely linked to adaptive traits or domestication. Finally, we investigate spatio-temporal gene expression patterns, with

particular focus on berry development and pathogen response to detect P450s with potential roles in these important physiological processes.

Material and methods

Gene annotation

We annotated the cytochromes P450 using the 12X.2 version of the assembly of the *Vitis vinifera* cv PN40024 genome ([9,36], https://urgi.versailles.inra.fr/Species/Vitis/Data-Sequences/Genome-sequences). Four publically available datasets of cytochromes P450 were used to perform similarity searches in the PN40024 genome. 947 protein sequences of grape P450s were downloaded from the NCBI Protein database (http://www.ncbi.nlm.nih.gov/protein, Feb 2014). Three datasets were downloaded from David Nelson's website (http://drnelson.uthsc.edu/CytochromeP450.html, Feb 2014), which stores manually curated annotations of cytochromes P450 for many species: 702 P450 protein sequences of *Vitis vinifera* cv Pinot Noir clone ENTAV115 (28, http://drnelson.uthsc.edu/vitis.htm); 416 P450 protein sequences of *Vitis vinifera* cv PN40024 from the 8x assembly version of the genome (10, http://drnelson.uthsc.edu/Vitis.additionalP450s.htm); and 288 P450 protein sequences of *Arabidopsis thaliana* (35, http://drnelson.uthsc.edu/Arabidopsis.Blast.file.html). Using these four datasets, we expected to be as exhaustive as possible in the cytochromes P450 similarly search of the PN40024 genome. The four datasets were masked for repeat sequences using the online tool "Repeat Masking" from Censor (http://www.girinst.org/censor/index.php).

The four masked datasets were used to perform four independent TBLASTN analyses [37] against the PN40024 12X.2 sequence with an e-value cutoff of 1e-3. The TBLASTN outputs were parsed using a homemade script. The hits from the three grape datasets were kept if they were at least 50 amino acids long with at least 70% sequence identity. The hits from the *Arabidopsis* dataset were kept if they were at least 50 amino acids long with an identity percentage of at least 50%. The software Exonerate (version 2.2.0, build October 2008, [38]) was used to predict gene structures using the protein2genome parameter and the same cutoff of sequence identity as above. A homemade script was used to reformat the output files from exonerate into files in the gff format. These gff files were imported to the Artemis genome browser [39] to perform the manual curation of the structures suggested by Exonerate. The parsed hits identified through TBLASTN were used to improve or to complete the Exonerate annotations. Every annotation starting with a start codon, ending with a stop codon and with correct exon-intron borders (GT-AG or sometimes GC-AG) was considered as a complete "gene". Every annotation showing this gene structure (start and stop codons, correct exon-intron borders) but with a single point mutation creating a frameshift, a premature stop codon or a wrong exon-intron border was considered as a "putative pseudogene" also marked "pseudogene?" because it may result from a mistake in the genome assembly. Every annotation interrupted by a gap in the genomic sequence or including one was considered as a "partial" annotation. All the other annotations with wrong gene structure but showing a significant similarity level with a cytochrome P450 from one of the four datasets were annotated as "pseudogenes". The genome annotation V1 stored in Grape Genome Database hosted at CRIBI ([40]; http://genomes.cribi.unipd.it/DATA/GFF/V1.phase.gff3) and a set of expertized and functional grape cytochromes P450 were used to guide the manual curation.

To validate the gene structure, two transcript datasets were used. First, the *Vitis vinifera* unigene set build #15 from the NCBI database was downloaded (ftp://ftp.ncbi.nih.gov/repository/UniGene/Vitis_vinifera/Vvi.seq.uniq.gz). The 32,193 unigenes were mapped on the PN40024 12X.2 sequence using GMAP version 2013-11-27 [41] using the default parameters except for

the format parameter which was set to "gff3_match_cdna". The second transcript dataset was locally assembled using six RNA-Seq experiments ([42], SRR519450, SRR519456, SRR520380 and SRR520385; [43], all four samples; [44], SRR493740- SRR493746; [45], SRR866544, SRR866570, SRR866571 and SRR866576; [46], SRR522472, SRR522477 and SRR522478; and four RNA-Seq datasets submitted in the frame of this study. The software Tophat2 v2.0.11 [47] was used to map the RNA-Seq reads against the PN40024 12X.2 sequence using the following parameters: -p 5 -N 5—read-edit-dist 5. The software Cufflinks v2.2.1 [48] was used to assemble the transcripts from all the RNA-Seq experiments. First the cufflinks command was used with the -p 5 parameter and then the cuffmerge command with the -p 15 parameter and using the fasta file of the PN40024 12X.2 sequence for the -s parameter. This assembly led to 32,219 transcripts and to a gtf file showing their mapped location in the PN40024 12X.2 sequence. The two transcript datasets were formatted in gff format compatible with the Artemis Browser so that the predicted gene structures of the cytochromes P450 could be compared with the transcripts and edited if needed.

The command maskFastaFromBed v2.19.1 from the bedtools package [49] was used to mask the regions of the PN40024 12X.2 sequence where we annotated cytochrome P450 exons after having reformatted the gff file of the annotations into a bed file. We performed TBLASTN analyses of the four grape cytochrome P450 datasets against the masked PN40024 12X.2 sequence and parsing analyses using the same parameters and cutoffs as previously described. This step allowed identifying the region of the grape genome for which a cytochrome P450 similarity was missed during the manual curation.

To validate the set of complete genes of cytochromes P450 that we annotated, a BLAST against non-redundant sequence database (NR) was performed and only the genes for which the best hit was a cytochrome P450 were kept. For the pseudogenes, a BLASTX was performed against the set of complete P450 genes that we annotated and we kept only the ones that aligned over at least 30% of the query length with the percentage identity of 50%.

The presence of physical clusters of cytochrome P450s in the grape genome was tested based on the following definition of a cluster. Two consecutive P450 annotations are part of a cluster if they are separated by 200kb and 8 non-P450 genes at the most [50,51]. The two annotations also have to be located on the same scaffold, which guaranties a precise estimation of the intergenic distances. A bootstrap test was performed to check whether the cytochromes P450 were more clustered than what is randomly expected. A homemade script was developed with R version 3.0.2 [52]. Ten thousand sampling without replacement of 579 (number of P450 annotations) or 279 features (number of complete P450 genes) were performed on the genome annotation V1 stored in Grape Genome Database hosted at CRIBI counting 29,971 features. The percentage of features organized in clusters was computed using the same protocol as for cytochromes P450. The p-value was calculated by counting each time a percentage equal of greater than the percentage of P450 in clusters divided by 10000 (number of iterations).

Sequence similarity within and between clusters was analyzed by performing a BLASTP search of translated complete P450 genes against themselves. Only the genes that aligned over at least 70% of the query length with the percentage identity of 40% were kept. The Circos software [53] was used to draw the figure. Clusters that contained less than two complete genes were excluded from this analysis (i.e. clusters that contained partial genes, pseudogenes and putative pseudogenes with less than 2 complete genes).

The dotter software version 4.23 [54] was used to draw the sequence similarity graphs of the cluster 190 with its fasta sequence and annotations in a gff format as an input.

Sequence classification

Cytochrome P450 genes, partial genes and putative pseudogenes were aligned to the P450 sequences from the heterozygous Pinot Noir genome, retrieved from the cytochrome P450 homepage (http://drnelson.uthsc.edu/CytochromeP450.html). In the case of protein sequence identity above 95%, the original name was kept. New sequences were assigned a family based on the best hit among already named grapevine P450s. Twenty-two sequences were given a new CYP name.

Phylogeny

Sequences from non-Vitis species were retrieved from the cytochrome P450 homepage (http://drnelson.uthsc.edu/CytochromeP450.html). Pseudogenes and incomplete genes were excluded from the analysis. 279 *Vitis vinifera* CYP (Fig 1) and 191 CYP76, 80 and 706 protein sequences from *Aquilegia caerulea*, *Nelumbo nucifera*, *Mimulus guttatus*, *Solanum lycopersicum*, *Amborella trichopoda*, *Oryza sativa*, *Brachypodium distachyon*, *Arabidopsis thaliana*, *Medicago trunculata*, *Populus trichocarpa* and *Vitis vinifera* (S1 Fig) were aligned with MUSCLE [55] implemented in Seaview [56,57]. Conserved sites were selected in the alignment using Gblocks [58] using the less stringent option parameters. Maximum likelihood phylogenies were obtained from the full-length alignments and from the subset of more conserved sites alignments (all *Vitis* CYP: 166 sites and 11 species CYP alignment: 278 sites) using RAxML (v 8.2.4) [59] via the CIPRES Science Gateway [60] and PhyML (implemented in Seaview v 4.5.4) [61]. Bootstrap values are shown on the nodes of the Vitis all CYP phylogeny. Nodes with bootstrap values below 60 were manually suppressed from the 11 species CYP phylogeny and are shown as trifurcations (unsolved topologies). The trees were visualized and colored using Figtree (http://tree.bio.ed.ac.uk/software/figtree). The species cladogram in (S1 Fig) was inferred from the APGIII system [62].

Gene expression

We retrieved raw grape RNA-Seq data from NCBI SRA public database (http://www.ncbi.nlm.nih.gov/sra). Fifty-nine sequence files generated in the framework of six different experiments [42,43,45,46,63,64] and four RNA-Seq datasets submitted in the frame of this study were used. The data were formatted in the fastq format using the fastq-dump command from the SRA Toolkit package version 2.3.4 (http://www.ncbi.nlm.nih.gov/books/NBK158900).

Alignments of these reads against the PN40024 12X.2 sequence were then performed using GSNAP version 2013-11-27 [65] with the following parameters: -B 4, -N 1, -n 3,—nofails and the quality protocol according to the experiment. These files were parsed to keep the best, unique and paired (if paired-end reads) alignments using a homemade script.

The number of fragments aligned on each annotation from the genome annotation V1 stored in Grape Genome Database hosted at CRIBI and the cytochromes P450 was counted using the command htseq-count from the HTSeq framework version 0.6.0 [66] with the following parameters: -m intersection-nonempty and -s no. Using a homemade script, FPKMs (Fragments Per Kilo base of exon per Million fragments mapped) were calculated for every annotation.

Using all non-zero FPKM values, the 33th and 66th quantiles were calculated to assign the expression values to one of the four levels of expression chosen: no, low, average and high expression. The experiments were grouped into six categories regarding the conditions under which the samples were obtained. These categories were: leaves, downy mildew (*Plasmopara viticola*) infected leaves, powdery mildew (*Erysiphe necator*) infected leaves, flowers, young berries and ripe berries. An average expression per category was then calculated for each gene

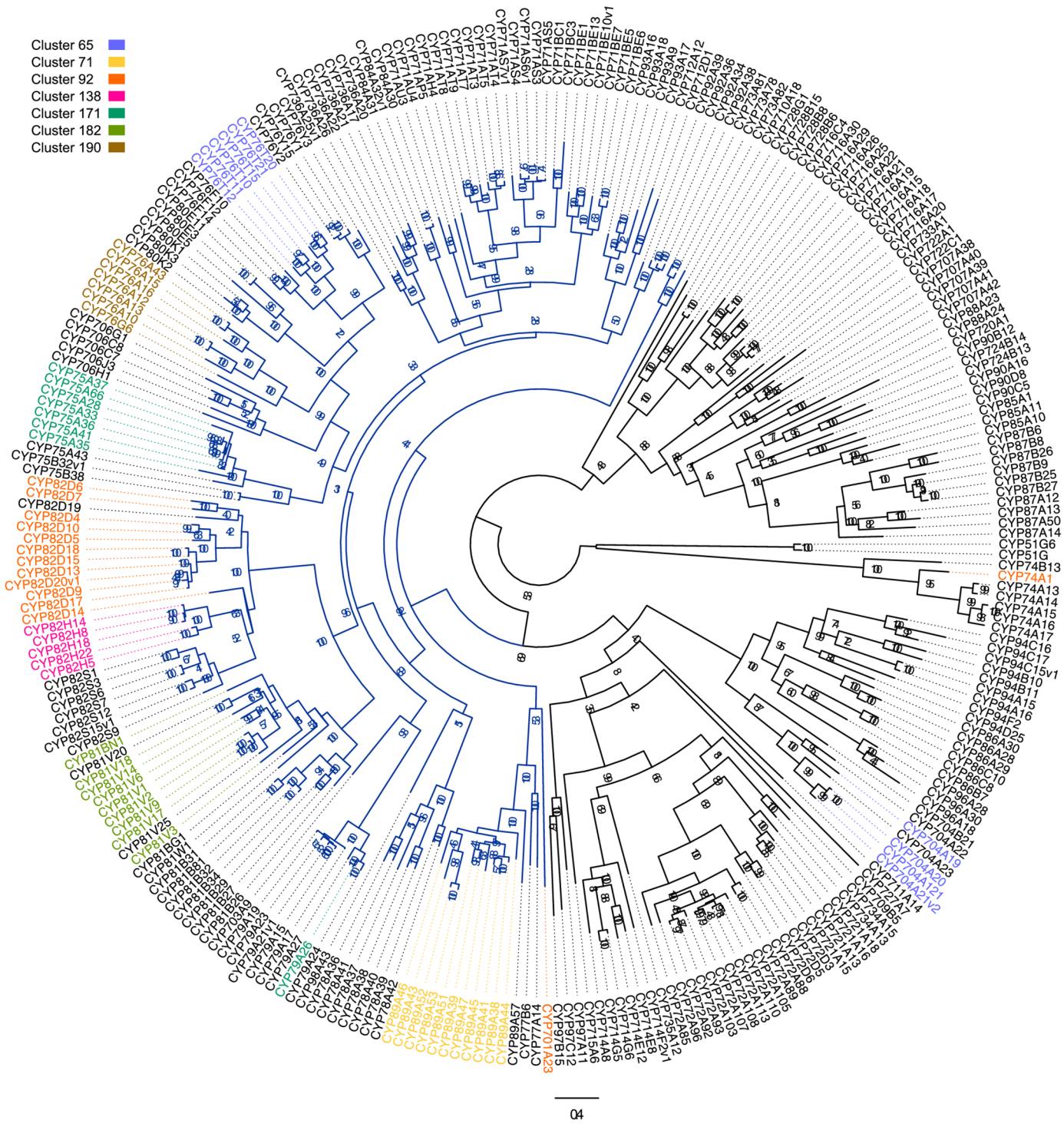

Fig 1. Molecular phylogenetic analysis of grapevine cytochrome P450. The alignment of full-length cytochrome P450 protein sequences was used to generate a maximum likelihood tree. The dark blue clade is the clan 71, which often contains genes involved in specialized metabolism. The highlighted genes belong to the seven largest physical clusters.

https://doi.org/10.1371/journal.pone.0199902.g001

and assigned to one of the four levels of expression regarding its value: no expression if the average was zero, low expression between zero and the 33% quantile, average expression between 33% and 66% quantile and high expression for averages higher than the 66% quantile.

The average expression values for each P450 annotation were used to perform a clustering analysis using HCE version 3.5 [67] with a complete linkage method and a Pearson's correlation as distance measure. The cut-off to define the clusters was set at a Pearson's correlation coefficient of 0.656. The heatmap was drawn using the package "pheatmap" [68] after row normalization ((FPKM value – row minimum) / row maximum) in R [52].

A RNA-Seq dataset of 48 conditions for berries at four developmental stages (bbch75 = pea size; bbch77 prior to veraison; bbch85 at the end of veraison; bbch89 ripe) for four grapevine varieties (Sangiovese, Barbera, Negro amaro and Refosco) in triplicate, published by Palumbo and coworkers [69] was used to perform an analysis of differentially expressed genes. The reads were aligned using STAR [70] and counted using featureCounts [71] on the grapevine reference genome PN40024 12X.2 and the VCost.v3 annotation [72] supplemented with the cytochrome P450 annotations. The analysis of differentially expressed genes across the four varieties was performed on the whole gene set of the grapevine genome using the script AskoR_DE.R (https://github.com/askomics/askoR) with the parameter cpm > 0.5.

Accession numbers

RNA-Seq datasets of green and mid-ripening berries from Riesling and Gewurztraminer were submitted at NCBI SRA public database under the BioProject accession number PRJNA378596. PRJNA254035, PRJNA168987, PRJNA244752, PRJNA203687 and PRJNA169607 RNA-Seq BioProjects were retrieved to complete the analysis.

Results

Gene annotation, classification and phylogeny

A similarity search of the *V*. *vinifera* PN40024 genome with known P450 sequences revealed 579 putative P450 sequences (S1 File). We manually curated the sequences obtained with a gene prediction algorithm, and validated the annotation with grapevine unigenes and RNAseq reads (see Material and methods). We distributed them into four categories: genes, partial genes, putative pseudogenes and pseudogenes. This led to the identification of 279 full-length genes, which is fewer than the 315 genes reported for the heterozygous Pinot Noir genome on the Cytochrome P450 homepage (http://drnelson.uthsc.edu/CytochromeP450.html), and suggests that some sequences previously annotated as different genes are probably allelic variants. The number of cytochromes P450 in grapevine is comparable to their number in other plants (e.g. 242 in *Arabidopsis thaliana*, 272 in *Solanum lycopersicum* and 309 in *Oryza sativa*). Twenty sequences were annotated as partial genes, lacking a segment of the sequence due to gaps in the genome assembly. Eleven putative pseudogenes only contain one nonsense mutation or frame shift, which could originate from sequencing errors or be genuine but still exist as functional genes in some varieties. Finally, the 269 pseudogenes are fragments, either containing multiple stop codons or frameshift mutations, or sequences not aligning to the whole length of homologous P450 genes.

Grapevine P450s can be assigned to 48 families based on sequence identities. A phylogenetic analysis either of the full-length sequences or of a subset of conserved P450 sites confirmed this classification for most of the families. One exception is CYP90B, which is clustered with CYP720 and CYP724 as previously observed [73]. Other exceptions are the families CYP76 and CYP80, which form a monophyletic group (Fig 1, see Material and methods). We thus investigated the phylogeny of these two families in the broader context of selected

angiosperm species (S1 Fig). CYP80 clearly groups with CYP76 sequences, but forms an independent clade between CYP76A/G and the rest of CYP76 sequences (labeled core CYP76). Within the CYP76A/G clade, a eudicot duplication gave rise to the two subfamilies CYP76A and CYP76G. Within the large "core CYP76" clade the uncertain position of both the monocot and *Amborella trichopoda* CYP76s could be due to a problem of long-branch attraction. A specific core eudicot duplication gave rise to CYP76F/B/X on one side and CYP76T/C/E on the other side. These tree topologies were obtained both with the full-length alignment and the partial alignment of conserved sites. Although species-specific "blooms" appeared in the whole CYP76/80 family, they are particularly abundant in the "core CYP76" clade. Different subfamilies expanded in different species.

Comparison of P450 family sizes between species (S2 Fig) allowed us to identify families that expanded in grapevine and might have a role in the production of species-specific specialized metabolites. An expansion of the CYP75 family, involved in anthocyanin biosynthesis, is already well documented [74], whereas the function of CYP82, the largest P450 family in grapevine with 25 members, is currently unknown in this species. Other families that are larger in grapevine than in most other species are: CYP76, CYP79, CYP80, CYP81, CYP87, CYP89 and CYP716.

Structural organization of the P450s in the PN40024 genome

The 579 cytochrome P450 sequences are distributed on all the 19 chromosomes. Some chromosomes, namely 18, 19 and 6, carry a large number of P450s (77, 57 and 51 sequences, respectively), whereas others, for example chromosome 5, carry very few (8 sequences) (S3 Fig). Twenty-four P450 sequences (7 genes, 6 partial genes, 11 pseudogenes) are located on the "Unknown" chromosome, which is composed of scaffolds that could not be anchored on any of the 19 chromosomes. Since the genome is not completely homozygous (estimated homozygosity is 93% [9]), the "Unknown" chromosome may also contain allelic variants of genes that are placed on the 19 chromosomes.

We further investigated the distribution of cytochrome P450 sequences in clusters or groups in close physical proximity (separated by less than 200 kb and eight non-P450 genes [50,51]). Our results show that P450 sequences are organized in clusters and not randomly distributed in the genome (bootstrap test, p-value < 0.0001). A large majority of cytochrome P450 sequences (452 or 78%) are part of one of the 85 clusters and only 22% (127 P450 sequences) are isolated in the grape genome. The largest number of clusters (40%) are only composed of two P450 sequences, whereas the largest cluster counts thirty-five P450 sequences. On average, there are five P450s per cluster and the median is three P450s per cluster (S4 Fig). The clusters are not enriched neither in complete genes nor pseudogenes, compared to isolated annotations (data not shown). Some chromosomes, such as 16 and 18, are enriched in clustered P450s, whereas others, such as chromosomes 4 and 11, are enriched in isolated P450 (Fig 2 and S3 Fig).

Cytochrome P450 families group genes with higher sequence similarity ($\geq$40% protein sequence identity) and often a similar function. A majority of physical clusters are composed of members of only one P450 family (63 clusters, 74%) and the remaining clusters are composed of up to three P450 families. The four largest clusters are composed of several P450 families, whereas the clusters with single P450 families are smaller (Fig 2 and S4 Fig). Most of the largest P450 families (CYP82, CYP71, CYP81, CYP76, CYP72 and others) are organized in clusters (S1 Table).

Clustering by P450 family already indicates that more similar P450 sequences cluster in closer physical proximity. But many P450 families are dispersed among several clusters. We

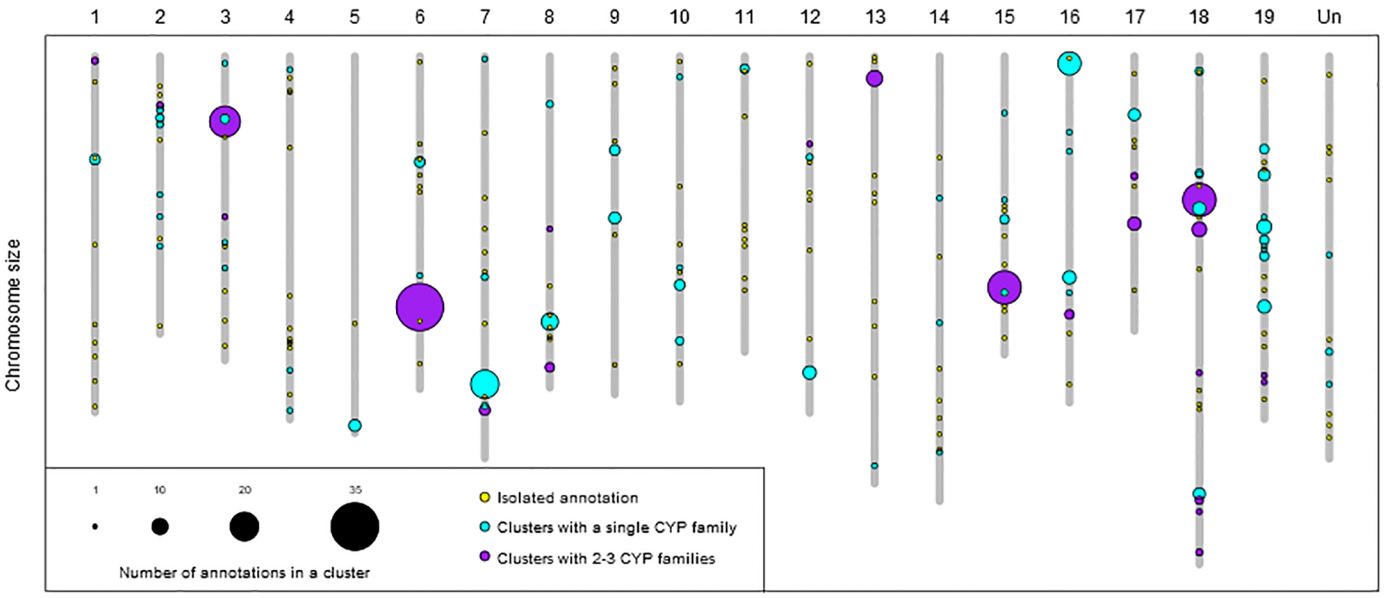

Fig 2. Physical map of cytochrome P450 sequences on the 19 V. vinifera chromosomes. Yellow circles represent isolated annotations, light blue circles represent physical clusters composed of members of only one P450 family and the purple circles represent physical clusters composed of members of 2–3 P450 families. The circle size is proportional to the number of sequences in the cluster. The numbers 1–19 are chromosome numbers and "Un" is "Unknown chromosome" which contains sequences with unknown chromosome location.

https://doi.org/10.1371/journal.pone.0199902.g002

thus wished to explore whether the closest paralogs belong to the same or different clusters (Fig 3). The majority of clustered P450 genes (86%) have their closest paralog (the best BLAST hit) in the same cluster. The second and third closest paralogs (second and third best BLAST hit) are in the same cluster for 58% and 49% of the clustered P450 genes. The sequence similarities within the same cluster are thus higher than between clusters.

Large P450 clusters in *V. vinifera* genome formed *via* different mechanisms

To investigate the mechanisms underlying the formation of large groups of physically close cytochrome P450 genes (hereafter called physical clusters), we further analyzed the sequence similarity within clusters, taking into account not only the coding P450 sequences, but also the surrounding non-coding-sequences. This allowed us to infer the mechanism of cluster formation. We focused on the seven largest physical clusters, which comprises between eleven to thirty-five P450 sequences (Table 1). Together, these seven clusters contain 23% of all P450 genes, and a similar fraction of total P450 sequences. Most of the sequences in these clusters are part of "clan 71", which is a large clade of plant cytochromes P450 often involved in the biosynthesis of species-specific adaptative metabolites (Fig 1).

Analysis of similarity blocks within these clusters showed they differ remarkably in their structures (S5 Fig). One of the largest physical clusters, cluster 65, is characterized by low similarities, both among the P450 sequences and surrounding non-coding regions. The similarity blocks of two other large physical clusters, 71 and 171, are restricted to P450 sequences and do not extend to the intergenic regions. Single gene duplications were thus probably the main mechanism of formation of these two clusters. The similarity blocks of physical clusters 138 and 182 extend to the non-coding regions around the cytochromes P450 annotations. This suggests the duplication events leading to formation of these clusters happened relatively recently. High similarity between the non-coding regions, which include the promoter regions, should result in similar expression profiles. Cluster 138 has the highest fraction (73%) of

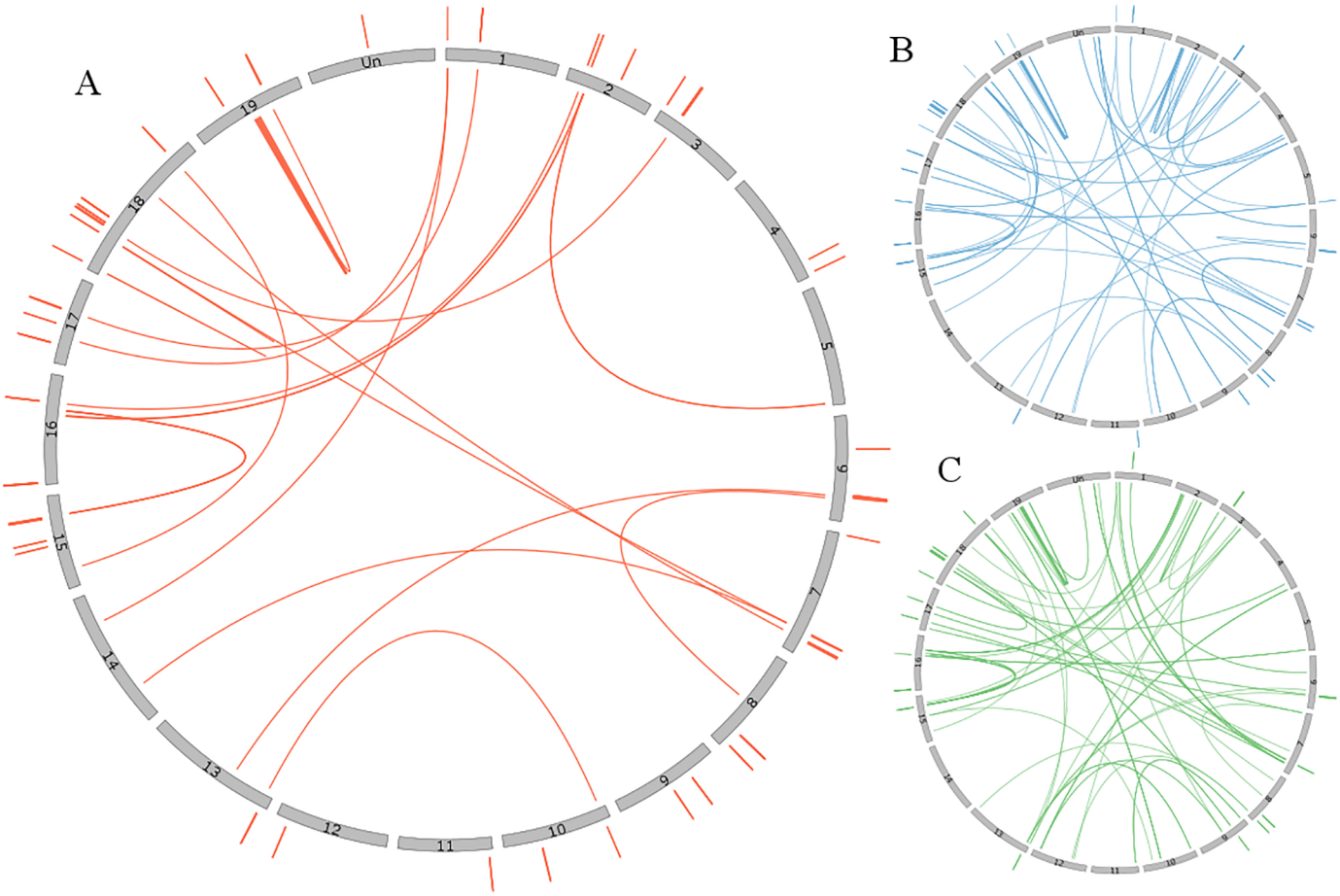

Fig 3. Similarity of the P450 genes between and within clusters. For each circle, the grey bars correspond to the 19 grape chromosomes and the "Unknown chromosome". The lines connect complete P450 genes according to their similarity. The lines outside the circles show the similarity between genes of the same cluster, whereas the lines in the circle connect similar genes of different clusters. Only P450 genes that form clusters composed of at least two complete genes are illustrated here. The seven largest clusters are labeled with numbers corresponding to Table 1. The lines are connecting the genes corresponding to the best BLAST hit (A), second best hit (B) or third best blast hit (C).

https://doi.org/10.1371/journal.pone.0199902.g003

pseudogenes of all the seven large clusters. In physical clusters 92 and 190, the similarity blocks extend over even longer regions that include three to four cytochrome P450 sequences and their intergenic regions (Fig 4). In addition, the type of annotation (gene or pseudogene) was also maintained in the same order between duplicated blocks. This suggests these two clusters formed through very recent proximal segmental duplications.

Expression profiles of grapevine P450s

To identify P450 genes with potential roles in pathogen resistance or biosynthesis of berry metabolites we analyzed the expression of the 579 P450 sequences. Pseudogenes were included in the analysis of expression to account for recently pseudogenized sequences that may still be expressed to some extent. We used 59 RNA-Seq datasets (S2 Table), which describe gene expression in different tissues (flowers, berries, leaves), different stages of berry development, and pathogen infection (downy and powdery mildews). To enable a meaningful comparison of gene expression between different experiments we calculated fragment number per kilobase of transcript per million mapped reads (FPKM) for each P450 sequence (S1 File). The majority

Table 1. Description of the seven largest physical P450 clusters in the *V. vinifera* genome.

Label	Chr	Location	Total seq.	Complete genes	Expressed seq.	Co-expression	Organization
65	15	15572751.. 15909327	20 CYP76 4 CYP704	10	20	Flowers and constitutive	Low similarity among members
71	16	401789.. 596606	16 CYP89	11	14	All leaves	Single gene duplications
92	18	9625486.. 9912876	22 CYP82 1 CYP74 1 CYP704	14	16	Powdery mildew infection and ripe berries	Duplicated blocks with co-expression; some single gene duplications
138	3	4387722.. 4512089	22 CYP82	5	16	Young berries	Small duplicated blocks, a few are co-expressed, single gene duplications
171	6	16790972.. 17446396	21 CYP75 14 CYP79	8	28	All leaves and powdery mildew infection	Single gene duplications
182	7	22260680.. 22372250	20 CYP81	9	15	Berries	Small duplicated blocks, a few are co-expressed, single gene duplications
190	8	18038159.. 18121816	11 CYP76	7	9	Flowers	Duplicated blocks with co-expression; some single gene duplication

Label—sequential number of each cluster in the genome; Chr—chromosome number; Location—chromosome coordinates; Total seq.—number of P450 sequences in each cluster, including complete and partial genes, putative pseudogenes and pseudogenes with their family distribution; Complete genes—number of complete P450 genes in the cluster; Expressed sequences—number of expressed P450 sequences in the cluster, co-expression—expression pattern of the cluster; Organization—description of structural organization and mechanism of formation of each cluster.

https://doi.org/10.1371/journal.pone.0199902.t001

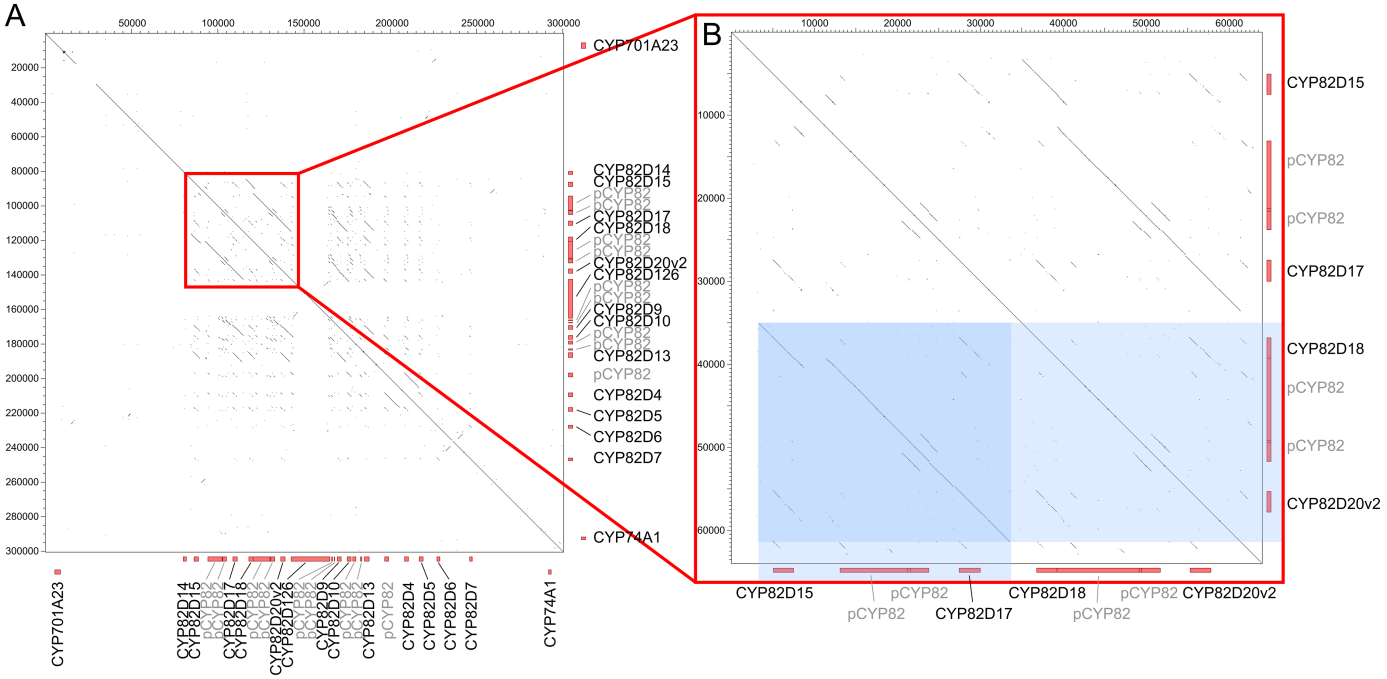

Fig 4. Dot matrix of segmental duplications in the physical cluster 92. Physical cluster 92 is located on chromosome 18 and comprises twenty-two CYP82 sequences, one CYP74 sequence and one CYP704 sequence. The dots and the black lines represent the sequence similarities in cluster **92** compared to itself. The red rectangles on the sides of the graph represent cytochrome P450 sequences. Complete genes are labeled with their name and pseudogenes are labeled with "p" and the P450 family. A) The similarities for the whole cluster **92**. B) A zoom of the red squared region, which contains two 30kb blocks with very high similarity. Analysis of gene expression showed that CYP82D15 and CYP82D18 are co-expressed (expression cluster **A**, expression in ripe berries) as well as CYP82D17 and CYP82D20v2 (expression cluster **C**, expression in downy mildew infected leaves). The pseudogenes of the enlarged segment are not expressed.

https://doi.org/10.1371/journal.pone.0199902.g004

of P450 sequences (457 or 79%) were expressed in at least two experiments. Of the remaining 122 non-expressed P450 sequences, only seven were complete genes. Out of the expressed P450 sequences, complete genes showed higher expression (mean FPKM = 10.6, median FPKM = 0.4) compared to pseudogenes and putative pseudogenes (mean FPKM = 2.4, median FPKM = 0) or partial genes (mean FPKM = 1.7, median = 0.1).

To simplify the dataset, we grouped the 59 experiments in six categories: leaves, downy mildew (*Plasmopara viticola*) infected leaves, powdery mildew (*Erysiphe necator*) infected leaves, flowers, young berries and ripe berries. "Flowers", "young" and "ripe berries" categories were significantly enriched in expressed P450 sequences whereas "leaves" and "downy mildew" categories were depleted (chi^2 test, p-value = 1E-15). In addition, we grouped the expression levels into four classes (no, low, medium or high expression). The "powdery mildew" category was found to be significantly enriched for highly expressed P450 sequences whereas the "leaves" category was depleted (chi^2 test, p-value = 9E-16). Altogether, these results indicate a significant induction of P450 expression caused by biotic stress, especially powdery mildew infection, and in grapevine organs synthesizing aromas and volatile compounds.

We analyzed the expression patterns by clustering the expression profiles of the 457 expressed cytochrome P450 sequences. A Pearsons' correlation coefficient cut-off of 0.656 resulted in eight expression clusters, shown in Fig 5. Only twenty-seven P450 sequences (6%, grouped in expression cluster G) are expressed constitutively, that means expressed at similar levels over the six categories, but the vast majority of the P450 genes are expressed in specific organs or under particular conditions. Among constitutively expressed genes, the CYP76 family was significantly enriched (chi^2 test, p-value = 3E-7).

A major shift in P450 expression pattern occurs during berry development. Indeed, the three largest expression clusters F (97 sequences), A (88 sequences) and H (76 sequences) grouped P450 sequences preferentially expressed in flowers, ripe berries and young berries, respectively. Among the large CYP families, CYP76, CYP716 and CYP714 were significantly enriched in cluster F whereas CYP72 was significantly depleted. CYP716 was also enriched in cluster H. CYP78 and CYP80 were significantly enriched in cluster A. In addition, this cluster featured the P450 gene with the highest expression in all experiments: CYP78A41 with an average FPKM value in ripe berries of 1292. Beside this expression analysis on multiple organs of the plant, an analysis of differentially expressed genes during berries development (four stages) for four grapevine varieties was performed using previously published data [70]. Two hundred and forty five P450 genes were significantly differentially expressed for at least one variety (S6 Fig). Notably, CYP76, CYP82, CYP71, CYP72, CYP75, CYP81 and CYP89 families that are enlarged in the grapevine genome and/or that are involved in secondary metabolism showed the greatest number of differentially expressed genes (20, 19, 19, 18, 18, 14 and 13, respectively). This result strengthens the hypothesis that these cytochrome P450 gene families play a major role in the biosynthesis of aroma compounds and especially the varietal-specific aromas.

The analysis of gene expression on multiple organs of grapevine also highlighted some P450 sequences with potential role in biotic stress response. Indeed, cluster D (59 sequences) and cluster C (35 sequences) grouped P450 sequences preferentially expressed during powdery and downy mildew infections, respectively. Among the large CYP families, CYP82, CYP87 and CYP736 were significantly enriched in cluster D whereas CYP72 was significantly enriched in cluster C.

We more thoroughly investigated co-expression of cytochrome P450 sequences within the physical clusters. We found that 245 out of the 343 expressed P450 sequences that are part of a physical cluster (71%) show an expression profile similar to at least one member of the same physical cluster. This level of co-expression is highly significant (bootstrap test, p-value$<$0.0001), which means that P450 physical clusters are highly enriched in co-expressed

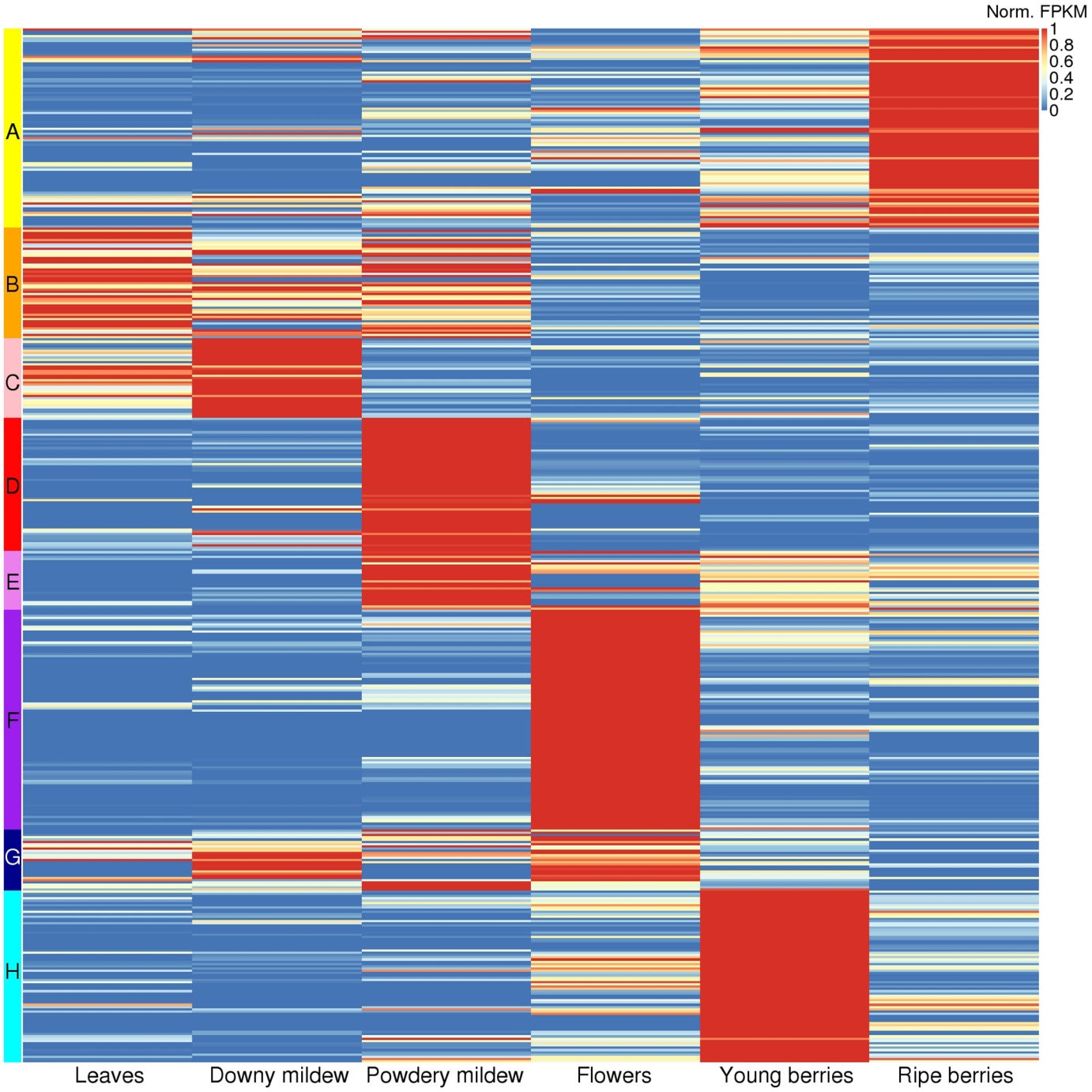

Fig 5. Heatmap of the P450 sequences, clustered according to their expression profile. The expression levels were averaged over the experiments classified in one of the six experimental categories: leaves, downy mildew (*Plasmopara viticola*) infected leaves, powdery mildew (*Erysiphe necator*) infected leaves, flowers, young berries and ripe berries. This heatmap includes the 457 expressed cytochrome P450 sequences. The color scale for the expression level represents FPKM values normalized by row ((FPKM value – row minimum) / row maximum). The color bars on the left are showing the eight expression clusters, which are designated by the letters on the side.

https://doi.org/10.1371/journal.pone.0199902.g005

sequences. The co-expression was analyzed further in the seven largest physical clusters to check whether tandem duplicated genes observed in the previous section maintained similar expression profiles (Table 1). Especially, we identified four large clusters with high similarity, not only among the coding, but also the non-coding regions. These non-coding regions

presumably include promoter sequences, so the genes in these clusters are expected to have the same expression pattern. Indeed, the 30 kb duplicated block within cluster 92 (Fig 4) retained the same expression profile after the segmental duplication. The duplicated segment consists of four P450 sequences: two genes and two pseudogenes. The two pseudogenes in both blocks were not expressed, whereas the two complete genes in both blocks (CYP82D15 and CYP82D18; CYP82D17 and CYP82D20v2) were co-expressed in ripe berries (cluster A) and downy mildew infected leaves (cluster C), respectively. Six out of sixteen expressed sequences in this cluster show induction in powdery mildew infected leaves (cluster D), whereas five other P450 sequences in the same cluster were up-regulated in ripe berries (cluster A). Interestingly, cluster 138 is also composed of CYP82 sequences, but these sequences were preferentially expressed in young berries (cluster H). Eight out of nine expressed CYP76 sequences in the physical cluster 190, on the other hand, were co-expressed in flowers (cluster F). In cluster 71, five out of fourteen expressed CYP89 genes were specifically expressed in leaves (cluster B). Cluster 182 grouped ten out of fifteen expressed CYP81 sequences specifically expressed in berries (5 sequences in cluster A and 5 sequences in cluster H).

Discussion

We produced an exhaustive, reliable and validated manual annotation of cytochromes P450 in the genome of the nearly homozygous grapevine (*V*. *vinifera)* accession PN40024 [9]. Cytochrome P450 superfamily in *Vitis vinifera* contains both very similar and very divergent genes (sequence identity ranges from 10% to almost 100%), and often form clusters in very close physical proximity, which makes it challenging for automated annotation algorithms. Manual curation is therefore necessary to produce a reliable annotation, suitable for demanding downstream applications such as phylogenetic or gene expression analysis. Grapevine P450s have been previously manually annotated (Cytochrome P450 homepage, http://drnelson.uthsc.edu/vitis.htm) in the highly heterozygous genome of Pinot noir cultivar [33]. Our annotation represents an improvement over the existing dataset for several reasons. The assembly of PN40024 genome is of better quality compared to the Pinot noir genome: it contains fewer gaps and a higher fraction of anchored contigs. The homozygosity of the genome not only enabled a better quality of the assembly, but also assured that most of the annotated sequences are individual loci and not allelic variants. This can partially explain a lower number of cytochrome P450 genes in our annotation—279—compared to the 315 genes reported on the Cytochrome P450 homepage. Additionally, the annotation on the Cytochrome P450 homepage classifies the sequences in only two categories, genes and pseudogenes, whereas we employed a more stringent classification into genes, partial genes, putative pseudogenes and pseudogenes. Lastly, we report the exact genomic coordinates of the P450 sequences, which facilitate comparison to annotations of other genes, and provide insights into structural organization of the grapevine CYPome.

Several gene families involved in the biosynthesis of specialized metabolites, such as terpene synthase and stilbene synthase genes, have expanded in grapevine genome compared to other species [9,75,76]. Although the total number of cytochrome P450 genes in grapevine is comparable to other species, individual P450 families experienced similar expansions. These expanded P450 families, similarly to terpene and stilbene synthases, form large physical clusters of more than ten homologous sequences. One of such families is CYP75, which together with CYP79 family members forms the largest physical cluster of thirty-five P450 sequences on chromosome 6. Expansion of CYP75 genes in grapevine was previously documented, but the presence of another P450 family, CYP79, in the same cluster was not reported [74]. Clustered genes with low or no homology sometimes participate in the same biosynthetic pathway

[77,78], but this is unlikely in the case of CYP75 and CYP79, since both families have well established roles in different biosynthetic pathways: CYP79 genes code for amino acid *N*-hydroxylases [27] for the synthesis of oxime derivatives precursors of cyanogenic glucosides, whereas CYP75A genes code for flavonoid 3',5'-hydroxylases [79], crucial enzymes in the biosynthesis of blue anthocyanins in the grape skin [74,80]. However, we cannot exclude the recruitment of some of these genes in other pathways. Interestingly, the sequencing of the genome of the grapevine cultivar Tannat, characterized by its very deep color, revealed an even higher number of CYP75 genes compared to the PN40024 accession [45]. Copy number of genes in a cluster can therefore vary between cultivars and could influence varietal characteristics. Other expanded P450 families in the grapevine genome are CYP82, CYP76, CYP81 and CYP89. They are forming large clusters resulting from gene duplication, a mechanism proven to favor emergence of new metabolic functions, especially for P450 genes [81].

Large-scale analysis of gene expression across several tissues and conditions provides a first hint to the putative P450 functions in grapevine. Pathogen infection causes a major shift in the P450 expression, inducing members from families CYP736, CYP81, CYP82 and CYP87. Their homologs in other species have been shown to participate in biosynthesis of highly specialized defense compounds (S1 Table). Interestingly, CYP736A25v1, which was shown to be upregulated upon infection with the Pierce disease pathogen *Xylella fastidiosa* [82], is also upregulated upon infection with powdery mildew and downy mildew pathogens. Two other sequences from the CYP736 family are also induced by biotic stress but their expression level is lower. Another large shift in expression occurs in developing grape berries. The most upregulated P450 families in the ripe berries expression cluster are CYP81, CYP82 and CYP78. Along with CYP76, CYP71, CYP72, CYP75 and CYP89, they are also differentially expressed across grapevine varieties during berry development. Therefore, these P450s are likely to participate in the biosynthesis of defense compounds or compounds important for the organoleptic properties of wine (aroma, colour, taste, mouthfeel).

The P450 gene with the overall highest expression level and the most up-regulated P450 gene in ripe berries is CYP78A41. A member of the same P450 family in tomato (*S. lycopersicum*) was selected during domestication to increase fruit size [24]. High expression of CYP78A41 in grape berries points to a similar event in grapevine domestication.

The most striking result in our analysis probably concerns the CYP76 family. From a structural point of view, we confirmed that CYP76 family is expanded in grapevine reference genome compared to most of the other plant genomes [35]. Furthermore, the CYP76 family was found to be highly clustered and part of the largest cluster, cluster190, in which a recent tandem duplication involving three consecutive genes was identified. This suggests an active evolution dynamics for this particular CYP family which could favor emergence of new metabolic functions [81] in particular grapevine varieties. We also found that the expression of the CYP76 family was mostly upregulated in grapevine flowers, similarly to what Boachon and coworkers [21] found in *Arabidopsis*. Interestingly, grapevine flowers, just like berries, are very rich in terpenes, some of the most important volatile compounds contributing to the floral bouquet of wine [75,83,84] and CYP76s were already found to be involved in terpene biosynthesis in *Arabidopsis* and grapevine [21,31,85,86]. More, CYP76 family showed the greatest number of genes with varietal-specific expression patterns during berry development. Altogether, its evolutionary dynamics along with its role in terpene pathway make the CYP76 genes major candidates to understand the diversity of grapevine and wine aromas, as recently shown for CYP76F14, a member of CYP76 family, which was identified as a key player in the production of wine lactone, a typical aroma of aged wines [25].

Conclusions

The phylogenetic and structural data suggest that some P450 families underwent multiple tandem or segmental duplications, which resulted in large physical clusters of homologous sequences that are often co-expressed. Most of these P450 families are involved in biosynthesis of highly specialized metabolites in other plant species. These genes are often expressed in specific conditions and tissues, such as leaves upon pathogen infection or during berry development. Finally, our work provides an exhaustive and robust annotation including clear identification, structural organization, evolutionary dynamics and expression patterns for the grapevine cytochrome P450 families, paving the way to efficient functional characterization of genes involved in grapevine defense pathways and aroma biosynthesis. Especially, our study points out towards the CYP76 family as the key candidate for further understanding the extraordinary diversity of grape and wine aromas.

Supporting information

S1 Fig. Phylogeny of CYP80 and CYP76 in angiosperms. Maximum likelihood tree of full length CYP76 and CYP80 protein sequences from a selection of angiosperms, rooted with CYP706 from all the included species. Nodes with bootstrap values below 60 are collapsed to trifurcations. Species specific clades with more than two members (except *V*. *vinifera*) are collapsed to triangles. The label of the triangle gives the subfamily and the number of members contained in the clade.
(PDF)

S2 Fig. Comparison of the number of P450 genes per family between species. Dot size is proportional to the relative family size (number of genes per family) in a given species compared to *Vitis vinifera* (Vv = *Vitis vinifera*, Nn = *Nelumbo nucifera*, Os = Oryza sativa, Bd = *Brachypodium distachyon*, Sl = *Solanum lycopersicum*, At = *Arabidopsis thaliana*, Pt = *Populus trichocarpa*, Gm = *Glycine max*, Mt = *Medicago truncatula*). The numbers in the first column are the absolute family sizes (numbers of genes per family) in *Vitis vinifera*. The number of genes per family was retrieved from the cytochrome P450 homepage. Pseudogenes and families not present in *V*. *vinifera* (CYP83, CYP99, CYP702, CYP705, CYP708, CYP718 and CYP729) were excluded from the count.
(PDF)

S3 Fig. Distribution of the *V*. *vinifera* P450s per chromosome. The blue bar corresponds to clustered annotations and the yellow bar to the isolated annotations. The "Unknown chromosome" is labeled as "Un".
(PDF)

S4 Fig. Distribution of the P450 sequences per physical cluster. Median and average values are labeled with arrows. The clusters composed of a single P450 family are represented in blue and those composed of 2 or 3 P450 families in orange.
(PDF)

S5 Fig. Dot matrix plots of the largest physical clusters. The dots and the black lines represent the sequence similarities in cluster 92 compared to itself. The red rectangles on the sides of the graph represent cytochrome P450 sequences. Complete genes are labeled with their name and pseudogenes are labeled with "p" and the P450 family.
(PDF)

S6 Fig. Heatmap of the differentially expressed cytochromes P450 between berries of four grapevine varieties. Expression in berries at four developmental stages (75 = pea size;

77 = prior to veraison; 85 = at the end of veraison; 89 = ripe) for four grapevine varieties (Sangiovese, Barbera, Negro amaro and Refosco) was studied. This heatmap shows the expression profiles of the 245 differentially expressed cytochromes P450 for at least one variety. The expression levels were averaged over the three replicates for each condition. The color scale for the expression level represents RPKM values normalized by row ((RPKM value – row minimum) / row maximum). The dendrogram on the left shows the clustering by gene. The raw data were obtained from [69].
(PDF)

S1 Table. List of P450 families with majority of its members grouped in physical clusters.
(PDF)

S2 Table. Description of RNA-Seq experiments used for analysis of gene expression.
(PDF)

S1 File. Fasta file of P450 cDNA, fasta file of P450 proteins, annotation file of P450s in gff3 format, FPKM of P450s in 59 experiments, RPKM for the analysis of differentially expressed P450 in the berries of four varieties, contrasts for the analysis of differentially expressed genes.
(XLSX)

Acknowledgments

We thank Adrian Arellano Davin for editing the CYP76 phylogenetic tree; Gisèle Butterlin and Lauriane Renault for technical assistance; and Fabrice Legeai for his technical support in using AskoR_DE.R. We also thank the two reviewers for their useful comments that helped to improve the manuscript.

Author Contributions

Conceptualization: Eric Duchêne, Frédéric Choulet, Etienne Paux, Philippe Hugueney, Danièle Werck-Reichhart, Camille Rustenholz.

Data curation: Tina Ilc, Gautier Arista, David R. Nelson, Camille Rustenholz.

Formal analysis: Tina Ilc, Gautier Arista, Raquel Tavares, Nicolas Navrot, Amandine Velt, Marc Fischer, Philippe Hugueney, Camille Rustenholz.

Funding acquisition: Philippe Hugueney, Danièle Werck-Reichhart.

Investigation: Tina Ilc, Gautier Arista, Raquel Tavares, Amandine Velt, Camille Rustenholz.

Methodology: Frédéric Choulet.

Project administration: Philippe Hugueney, Danièle Werck-Reichhart, Camille Rustenholz.

Supervision: Philippe Hugueney, Danièle Werck-Reichhart, Camille Rustenholz.

Validation: Tina Ilc, David R. Nelson, Camille Rustenholz.

Visualization: Tina Ilc, Raquel Tavares.

Writing – original draft: Tina Ilc, Gautier Arista, Raquel Tavares, Nicolas Navrot, Camille Rustenholz.

Writing – review & editing: Tina Ilc, Gautier Arista, Raquel Tavares, Nicolas Navrot, Eric Duchêne, Frédéric Choulet, Etienne Paux, Marc Fischer, David R. Nelson, Philippe Hugueney, Danièle Werck-Reichhart, Camille Rustenholz.

References

1. Zohary D, Spiegel-Roy P. Beginnings of Fruit Growing in the Old World. Science (80-). 1972; 187: 319–327.
2. FAOSTAT. Food and Agriculture Organisation fo the United Nations. Food and Agricultural commodities production / Commodities by region [Internet]. 2015 [cited 1 Jan 2015]. http://www.fao.org/faostat/en/#home
3. Myles S, Boyko AR, Owens CL, Brown PJ, Grassi F, Aradhya MK, et al. Genetic structure and domestication history of the grape. Proc Natl Acad Sci U S A. 2011; 108: 3530–3535. https://doi.org/10.1073/pnas.1009363108 PMID: 21245334
4. Duchêne E, Schneider C. Grapevine and climatic changes: a glance at the situation in Alsace. Agron Sustain Dev. 2005; 25: 93–99. https://doi.org/10.1051/agro:2004057
5. Duchêne E, Huard F, Dumas V, Schneider C, Merdinoglu D. The challenge of adapting grapevine varieties to climate change. Clim Res. 2010; 41: 193–204. https://doi.org/10.3354/cr00850
6. Bisson LF, Waterhouse AL, Ebeler SE, Walker MA, Lapsley JT. The present and future of the international wine industry. Nature. 2002; 418: 696–699. https://doi.org/10.1038/nature01018 PMID: 12167877
7. Borneman AR, Schmidt SA, Pretorius IS. At the cutting-edge of grape and wine biotechnology. Trends Genet. Elsevier Ltd; 2013; 29: 263–271. https://doi.org/10.1016/j.tig.2012.10.014 PMID: 23218459
8. Gray DJ, Li ZT, Dhekney SA. Precision breeding of grapevine (Vitis vinifera L.) for improved traits. Plant Sci. 2014; 228: 3–10. https://doi.org/10.1016/j.plantsci.2014.03.023 PMID: 25438781
9. Jaillon O, Aury J-M, Noel B, Policriti A, Clepet C, Casagrande A, et al. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature. 2007; 449: 463–7. https://doi.org/10.1038/nature06148 PMID: 17721507
10. Langridge P, Fleury D. Making the most of "omics" for crop breeding. Trends Biotechnol. 2011; 29: 33–40. https://doi.org/10.1016/j.tibtech.2010.09.006 PMID: 21030098
11. Nelson D, Werck-Reichhart D. A P450-centric view of plant evolution. Plant J. 2011; 66: 194–211. https://doi.org/10.1111/j.1365-313X.2011.04529.x PMID: 21443632
12. Bak S, Beisson F, Bishop G, Hamberger B, Höfer R, Paquette S, et al. Cytochromes P450. Arab B. American Society of Plant Biologists; 2011; 9: e0144. https://doi.org/10.1199/tab.0144 PMID: 22303269
13. Helliwell CA, Sheldon CC, Olive MR, Walker AR, Zeevaart JAD, Peacock WJ, et al. Cloning of the Arabidopsis ent-kaurene oxidase gene GA3. Proc Natl Acad Sci. 1998; 95: 9019–9024. https://doi.org/10.1073/pnas.95.15.9019 PMID: 9671797
14. Booker J, Sieberer T, Wright W, Williamson L, Willett B, Stirnberg P, et al. MAX1 Encodes a Cytochrome P450 Family Member that Acts Downstream of MAX3/4 to Produce a Carotenoid-Derived Branch-Inhibiting Hormone. Dev Cell. 2005; 8: 443–449. https://doi.org/10.1016/j.devcel.2005.01.009 PMID: 15737939
15. Anastasiou E, Kenz S, Gerstung M, MacLean D, Timmer J, Fleck C, et al. Control of Plant Organ Size by KLUH/CYP78A5-Dependent Intercellular Signaling. Dev Cell. 2007; 13: 843–856. https://doi.org/10.1016/j.devcel.2007.10.001 PMID: 18061566
16. Liu Z, Boachon B, Lugan R, Tavares R, Erhardt M, Mutterer J, et al. A Conserved Cytochrome P450 Evolved in Seed Plants Regulates Flower Maturation. Mol Plant. Elsevier; 2015; 8: 1751–1765. https://doi.org/10.1016/j.molp.2015.09.002 PMID: 26388305
17. Ito T, Meyerowitz EM. Overexpression of a gene encoding a cytochrome P450, CYP78A9, induces large and seedless fruit in arabidopsis. Plant Cell. 2000; 12: 1541–50. PMID: 11006330
18. Ehlting J, Hamberger B, Million-Rousseau R, Werck-Reichhart D. Cytochromes P450 in phenolic metabolism. Phytochem Rev. 2006; 5: 239–270. https://doi.org/10.1007/s11101-006-9025-1
19. Wellesen K, Durst F, Pinot F, Benveniste I, Nettesheim K, Wisman E, et al. Functional analysis of the LACERATA gene of Arabidopsis provides evidence for different roles of fatty acid ω-hydroxylation in development. Proc Natl Acad Sci U S A. 2001; 98: 9694–9699. https://doi.org/10.1073/pnas.171285998 PMID: 11493698
20. Lee S, Badieyan S, Bevan DR, Herde M, Gatz C, Tholl D. Herbivore-induced and floral homoterpene volatiles are biosynthesized by a single P450 enzyme (CYP82G1) in Arabidopsis. Proc Natl Acad Sci U S A. 2010; 107: 21205–10. https://doi.org/10.1073/pnas.1009975107 PMID: 21088219

21. Boachon B, Junker RR, Miesch L, Bassard J-E, Höfer R, Caillieaudeaux R, et al. CYP76C1 (Cytochrome P450)-Mediated Linalool Metabolism and the Formation of Volatile and Soluble Linalool Oxides in Arabidopsis Flowers: A Strategy for Defense against Floral Antagonists. Plant Cell. 2015; 27: 2972–90. https://doi.org/10.1105/tpc.15.00399 PMID: 26475865

22. Nafisi M, Goregaoker S, Botanga CJ, Glawischnig E, Olsen CE, Halkier B a, et al. Arabidopsis cytochrome P450 monooxygenase 71A13 catalyzes the conversion of indole-3-acetaldoxime in camalexin synthesis. Plant Cell. 2007; 19: 2039–52. https://doi.org/10.1105/tpc.107.051383 PMID: 17573535

23. Hansen CH. Cytochrome P450 CYP79F1 from Arabidopsis Catalyzes the Conversion of Dihomomethionine and Trihomomethionine to the Corresponding Aldoximes in the Biosynthesis of Aliphatic Glucosinolates. J Biol Chem. 2001; 276: 11078–11085. https://doi.org/10.1074/jbc.M010123200 PMID: 11133994

24. Chakrabarti M, Zhang N, Sauvage C, Muños S, Blanca J, Cañizares J, et al. A cytochrome P450 regulates a domestication trait in cultivated tomato. Proc Natl Acad Sci U S A. 2013; 110: 17125–30. https://doi.org/10.1073/pnas.1307313110 PMID: 24082112

25. Ilc T, Halter D, Miesch L, Lauvoisard F, Kriegshauser L, Ilg A, et al. A grapevine cytochrome P450 generates the precursor of wine lactone, a key odorant in wine. New Phytol. 2017; 213: 264–274. https://doi.org/10.1111/nph.14139 PMID: 27560385

26. Gómez JF, Talle B, Wilson ZA. Anther and pollen development: A conserved developmental pathway. J Integr Plant Biol. 2015; 57: 876–891. https://doi.org/10.1111/jipb.12425 PMID: 26310290

27. Hamberger B, Bak S. Plant P450s as versatile drivers for evolution of species-specific chemical diversity. Philos Trans R Soc Lond B Biol Sci. 2013; 368: 20120426. https://doi.org/10.1098/rstb.2012.0426 PMID: 23297350

28. Feyereisen R. Arthropod CYPomes illustrate the tempo and mode in P450 evolution. Biochim Biophys Acta—Proteins Proteomics. Elsevier B.V.; 2011; 1814: 19–28. https://doi.org/10.1016/j.bbapap.2010.06.012 PMID: 20601227

29. Swaminathan S, Morrone D, Wang Q, Fulton DB, Peters RJ. CYP76M7 is an ent-cassadiene C11alpha-hydroxylase defining a second multifunctional diterpenoid biosynthetic gene cluster in rice. Plant Cell. 2009; 21: 3315–25. https://doi.org/10.1105/tpc.108.063677 PMID: 19825834

30. Wang Q, Hillwig ML, Okada K, Yamazaki K, Wu Y, Swaminathan S, et al. Characterization of CYP76M5-8 indicates metabolic plasticity within a plant biosynthetic gene cluster. J Biol Chem. 2012; 287: 6159–68. https://doi.org/10.1074/jbc.M111.305599 PMID: 22215681

31. Hofer R, Boachon B, Renault H, Gavira C, Miesch L, Iglesias J, et al. Dual function of the cytochrome P450 CYP76 family from Arabidopsis thaliana in the metabolism of monoterpenols and phenylurea herbicides. Plant Physiol. 2014; 166: 1149–1161. https://doi.org/10.1104/pp.114.244814 PMID: 25082892

32. Miettinen K, Dong L, Navrot N, Schneider T, Burlat V, Pollier J, et al. The seco-iridoid pathway from Catharanthus roseus. Nat Commun. 2014; 5: 3606. https://doi.org/10.1038/ncomms4606 PMID: 24710322

33. Velasco R, Zharkikh A, Troggio M, Cartwright D a, Cestaro A, Pruss D, et al. A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. PLoS One. 2007; 2: e1326. https://doi.org/10.1371/journal.pone.0001326 PMID: 18094749

34. Nelson DR. The cytochrome P450 homepage. Hum Genomics. 2009; 4: 59–65. PMID: 19951895

35. Nelson DR, Ming R, Alam M, Schuler MA. Comparison of Cytochrome P450 Genes from Six Plant Genomes. Trop Plant Biol. 2008; 1: 216–235. https://doi.org/10.1007/s12042-008-9022-1

36. Grimplet J, Van Hemert J, Carbonell-Bejerano P, Díaz-Riquelme J, Dickerson J, Fennell A, et al. Comparative analysis of grapevine whole-genome gene predictions, functional annotation, categorization and integration of the predicted gene sequences. BMC Res Notes. 2012; 5: 213. https://doi.org/10.1186/1756-0500-5-213 PMID: 22554261

37. Altschul SF, Madden TL, Schaffer AA, Zhang JH, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997; 25: 3389–3402. PMID: 9254694

38. The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature. 2000; 408: 796–814. https://doi.org/10.1038/35048692 PMID: 11130711

39. Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, et al. Artemis: sequence visualization and annotation. Bioinformatics. 2000; 16: 944–945. Available: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=11120685 PMID: 11120685

40. Vitulo N, Forcato C, Carpinelli E, Telatin A, Campagna D, D'Angelo M, et al. A deep survey of alternative splicing in grape reveals changes in the splicing machinery related to tissue, stress condition and genotype. BMC Plant Biol. 2014; 14: 99. https://doi.org/10.1186/1471-2229-14-99 PMID: 24739459

41. Wu TD, Watanabe CK. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics. 2005; 21: 1859–75. https://doi.org/10.1093/bioinformatics/bti310 PMID: 15728110

42. Perazzolli M, Moretto M, Fontana P, Ferrarini A, Velasco R, Moser C, et al. Downy mildew resistance induced by Trichoderma harzianum T39 in susceptible grapevines partially mimics transcriptional changes of resistant genotypes. BMC Genomics. BMC Genomics; 2012; 13: 1.

43. Sweetman C, Wong DC, Ford CM, Drew DP. Transcriptome analysis at four developmental stages of grape berry (Vitis vinifera cv. Shiraz) provides insights into regulated and coordinated gene expression. BMC Genomics. BMC Genomics; 2012; 13: 691. https://doi.org/10.1186/1471-2164-13-691 PMID: 23227855

44. Vannozzi A, Dry IB, Fasoli M, Zenoni S, Lucchin M. Genome-wide analysis of the grapevine stilbene synthase multigenic family: genomic organization and expression profiles upon biotic and abiotic stresses. BMC Plant Biol. BMC Plant Biology; 2012; 12: 1.

45. Da Silva C, Zamperin G, Ferrarini A, Minio A, Molin D, Venturini L, et al. The High Polyphenol Content of Grapevine Cultivar Tannat Berries Is Conferred Primarily by Genes That Are Not Shared with the Reference Genome. 2013; 1: 1–13. https://doi.org/10.1105/tpc.113.118810 PMID: 24319081

46. Venturini L, Ferrarini A, Zenoni S, Tornielli GB, Fasoli M, Santo SD, et al. De novo transcriptome characterization of Vitis vinifera cv. Corvina unveils varietal diversity. BMC Genomics. BMC Genomics; 2013; 14: 1.

47. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 2013; 14: R36. https://doi.org/10.1186/gb-2013-14-4-r36 PMID: 23618408

48. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. Nature Publishing Group; 2010; 28: 511–5. https://doi.org/10.1038/nbt.1621 PMID: 20436464

49. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010; 26: 841–2. https://doi.org/10.1093/bioinformatics/btq033 PMID: 20110278

50. Richly E, Kurth J, Leister D. Mode of amplification and reorganization of resistance genes during recent Arabidopsis thaliana evolution. Mol Biol Evol. 2002; 19: 76–84. https://doi.org/10.1093/oxfordjournals.molbev.a003984 PMID: 11752192

51. Yang S, Zhang X, Yue J-X, Tian D, Chen J-Q. Recent duplications dominate NBS-encoding gene expansion in two woody species. Mol Genet Genomics. 2008; 280: 187–198. https://doi.org/10.1007/s00438-008-0355-0 PMID: 18563445

52. R Core Team. R: A Language and Environment for Statistical Computing [Internet]. 2016. https://www.r-project.org/

53. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, et al. Circos: an information aesthetic for comparative genomics. Genome Res. 2009; 19: 1639–45. https://doi.org/10.1101/gr.092759.109 PMID: 19541911

54. Sonnhammer ELL, Durbin R. A Dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. Gene. 1995; 167: 1–10.

55. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004; 32: 1792–1797. https://doi.org/10.1093/nar/gkh340 PMID: 15034147

56. Galtier N, Gouy M, Gautier C. SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. Bioinformatics. 1996; 12: 543–548. https://doi.org/10.1093/bioinformatics/12.6.543

57. Gouy M, Guindon S, Gascuel O. SeaView Version 4: A Multiplatform Graphical User Interface for Sequence Alignment and Phylogenetic Tree Building. Mol Biol Evol. 2010; 27: 221–224. https://doi.org/10.1093/molbev/msp259 PMID: 19854763

58. Castresana J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol. 2000; 17: 540–552. https://doi.org/10.1093/oxfordjournals.molbev.a026334 PMID: 10742046

59. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014; 30: 1312–1313. https://doi.org/10.1093/bioinformatics/btu033 PMID: 24451623

60. Miller MA, Pfeiffer W, Schwartz T. Creating the CIPRES Science Gateway for inference of large phylogenetic trees. 2010 Gateway Computing Environments Workshop (GCE). IEEE; 2010. pp. 1–8. 10.1109/GCE.2010.5676129

61. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. Syst Biol. 2010; 59: 307–321. https://doi.org/10.1093/sysbio/syq010 PMID: 20525638

62. The angiosperm phylogeny group. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG III. Bot J Linn Soc. 2009; 161: 105–121. https://doi.org/10.1111/j.1095-8339.2009.00996.x

63. Jones L, Riaz S, Morales-Cruz A, Amrine KCH, McGuire B, Gubler WD, et al. Adaptive genomic structural variation in the grape powdery mildew pathogen, Erysiphe necator. BMC Genomics. 2014; 15: 1081. https://doi.org/10.1186/1471-2164-15-1081 PMID: 25487071

64. Ramos MJ, Coito J, Silva H, Cunha J, Costa MM, Rocheta M. Flower development and sex specification in wild grapevine. BMC Genomics. 2014; 15: 1095. https://doi.org/10.1186/1471-2164-15-1095 PMID: 25495781

65. Wu TD, Nacu S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics. 2010; 26: 873–881. https://doi.org/10.1093/bioinformatics/btq057 PMID: 20147302

66. Anders S, Pyl PT, Huber W. HTSeq—A Python framework to work with high-throughput sequencing data. Bioinformatics. 2014; 31: 166–169. https://doi.org/10.1093/bioinformatics/btu638 PMID: 25260700

67. Seo J, Gordish-Dressman H, Hoffman EP. An interactive power analysis tool for microarray hypothesis testing and generation. Bioinformatics. 2006; 22: 808–14. https://doi.org/10.1093/bioinformatics/btk052 PMID: 16418236

68. Kolde R. pheatmap: Pretty Heatmaps [Internet]. 2015. https://cran.r-project.org/package=pheatmap

69. Palumbo MC, Zenoni S, Fasoli M, Massonnet M, Farina L, Castiglione F, et al. Integrated Network Analysis Identifies Fight-Club Nodes as a Class of Hubs Encompassing Key Putative Switch Genes That Induce Major Transcriptome Reprogramming during Grapevine Development. Plant Cell Online. 2014; 26. https://doi.org/10.1105/tpc.114.133710 PMID: 25490918

70. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: Ultrafast universal RNA-seq aligner. Bioinformatics. 2013; 29: 15–21. https://doi.org/10.1093/bioinformatics/bts635 PMID: 23104886

71. Liao Y, Smyth GK, Shi W. FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014; 30: 923–930. https://doi.org/10.1093/bioinformatics/btt656 PMID: 24227677

72. Canaguier A, Grimplet J, Di Gaspero G, Scalabrin S, Duchêne E, Choisne N, et al. A new version of the grapevine reference genome assembly (12X.v2) and of its annotation (VCost.v3). Genomics Data. 2017; 14. https://doi.org/10.1016/j.gdata.2017.09.002 PMID: 28971018

73. Nelson DR, Schuler M a, Paquette SM, Werck-Reichhart D, Bak S. Comparative genomics of rice and Arabidopsis. Analysis of 727 cytochrome P450 genes and pseudogenes from a monocot and a dicot. Plant Physiol. 2004; 135: 756–772. https://doi.org/10.1104/pp.104.039826 PMID: 15208422

74. Falginella L, Castellarin SD, Testolin R, Gambetta G a, Morgante M, Di Gaspero G. Expansion and subfunctionalisation of flavonoid 3',5'-hydroxylases in the grapevine lineage. BMC Genomics. BioMed Central Ltd; 2010; 11: 562. https://doi.org/10.1186/1471-2164-11-562 PMID: 20939908

75. Martin DM, Aubourg S, Schouwey MB, Daviet L, Schalk M, Toub O, et al. Functional Annotation, Genome Organization and Phylogeny of the Grapevine (Vitis vinifera) Terpene Synthase Gene Family Based on Genome Assembly, FLcDNA Cloning, and Enzyme Assays. BMC Plant Biol. BioMed Central Ltd; 2010; 10: 226. https://doi.org/10.1186/1471-2229-10-226 PMID: 20964856

76. Parage C, Tavares R, Réty S, Baltenweck-guyot R, Poutaraud A, Renault L, et al. Structural, Functional, and Evolutionary Analysis of the Unusually Large Stilbene Synthase Gene Family in Grapevine 1 [W]. 2012; 160: 1407–1419.

77. Takos AM, Rook F. Why biosynthetic genes for chemical defense compounds cluster. Trends Plant Sci. Elsevier Ltd; 2012; 17: 383–388. https://doi.org/10.1016/j.tplants.2012.04.004 PMID: 22609284

78. Nützmann H-W, Osbourn A. Gene clustering in plant specialized metabolism. Curr Opin Biotechnol. 2014; 26: 91–9. https://doi.org/10.1016/j.copbio.2013.10.009 PMID: 24679264

79. Ayabe S, Akashi T. Cytochrome P450s in flavonoid metabolism. Phytochem Rev. 2006; 5: 271–282. https://doi.org/10.1007/s11101-006-9007-3

80. Falginella L, Di Gaspero G, Castellarin SD. Expression of flavonoid genes in the red grape berry of "Alicante Bouschet" varies with the histological distribution of anthocyanins and their chemical composition. Planta. 2012; 236: 1037–51. https://doi.org/10.1007/s00425-012-1658-2 PMID: 22552639

81. Liu Z, Tavares R, Forsythe ES, André F, Lugan R, Jonasson G, et al. Evolutionary interplay between sister cytochrome P450 genes shapes plasticity in plant metabolism. Nat Commun. 2016; 7: 13026. https://doi.org/10.1038/ncomms13026 PMID: 27713409

82. Cheng DW, Lin H, Takahashi Y, Walker MA, Civerolo EL, Stenger DC. Transcriptional regulation of the grape cytochrome P450 monooxygenase gene CYP736B expression in response to Xylella fastidiosa infection. BMC Plant Biol. 2010; 10: 135. https://doi.org/10.1186/1471-2229-10-135 PMID: 20591199

83. Martin DM, Toub O, Chiang A, Lo BC, Ohse S, Lund ST, et al. The bouquet of grapevine (Vitis vinifera L. cv. Cabernet Sauvignon) flowers arises from the biosynthesis of sesquiterpene volatiles in pollen grains. Proc Natl Acad Sci. 2009; 106: 7245–7250. https://doi.org/10.1073/pnas.0901387106 PMID: 19359488

84. Lund ST, Bohlmann J. The Molecular Basis for Wine Grape Quality-A Volatile Subject. Science (80-). 2006; 311.

85. Ginglinger J-F, Boachon B, Höfer R, Paetz C, Köllner TG, Miesch L, et al. Gene Coexpression Analysis Reveals Complex Metabolism of the Monoterpene Alcohol Linalool in Arabidopsis Flowers. Plant Cell Online. 2013;25. Available: http://www.plantcell.org/content/25/11/4640.long

86. Costantini L, Kappel CD, Trenti M, Battilana J, Emanuelli F, Sordo M, et al. Drawing Links from Transcriptome to Metabolites: The Evolution of Aroma in the Ripening Berry of Moscato Bianco (Vitis vinifera L.). Front Plant Sci. Frontiers Media SA; 2017; 8: 780. https://doi.org/10.3389/fpls.2017.00780 PMID: 28559906